应用光学
Journal of Applied Optics
SCOPUS,CSCD等来源期刊
中文核心期刊、中国科技核心期刊

热成像特征中期融合夜视密集人群计数

任国印 吕晓琪 李宇豪

Night vision dense crowd counting based on mid-term fusion of thermal imaging features

REN Guoyin, LYU Xiaoqi, LI Yuhao

引用本文:
任国印, 吕晓琪, 李宇豪. 热成像特征中期融合夜视密集人群计数[J]. 应用光学, 2022, 43(6): 1088-1096. DOI: 10.5768/JAO202243.0604007
REN Guoyin, LYU Xiaoqi, LI Yuhao. Night vision dense crowd counting based on mid-term fusion of thermal imaging features[J]. Journal of Applied Optics, 2022, 43(6): 1088-1096. DOI: 10.5768/JAO202243.0604007

在线阅读 View online: https://doi.org/10.5768/JAO202243.0604007

文章编号：1002-2082 (2022) 06-1088-09

热成像特征中期融合夜视密集人群计数

任国印，吕晓琪，李宇豪

（内蒙古科技大学 机械工程学院，内蒙古 包头 014010）

摘　要：为了提高人群计数模型对尺度和光噪声的鲁棒性，设计了一种多模态图像融合网络。提出了一种针对夜间人群统计模型，并设计了一个子网络 Rgb-T-net，网络融合了热成像特征和可见光图像的特征，增强了网络对热成像和夜间人群特征的判断能力。模型采用自适应高斯核对密度图进行回归，在 Rgb-T-CC 数据集上完成了夜视训练和测试。经验证网络平均绝对误差为 18.16，均方误差为 32.14，目标检测召回率为 97.65%，计数性能和检测表现优于当前最先进的双峰融合方法。实验结果表明，所提出的多模态特征融合网络能够解决夜视环境下的计数与检测问题，消融实验进一步证明了融合模型各部分参数的有效性。

关键词：夜视环境人群计数；Rgb 图像；热成像；Rgb-T 特征融合

中图分类号：TN223　　**文献标志码：**A　　**DOI：**10.5768/JAO202243.0604007

Night vision dense crowd counting based on mid-term fusion of thermal imaging features

REN Guoyin, LYU Xiaoqi, LI Yuhao

（School of Mechanical Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China）

Abstract: In order to improve the robustness of crowd counting model to scale and optical noise, a multimodal image fusion network was designed. A statistical model for night crowd was proposed, and a sub network Rgb-T-net was designed. The network integrated the characteristics of thermal imaging and visible image, and the ability of network to judge the characteristics of thermal imaging and night crowd was enhanced. The proposed model used the adaptive Gaussian checking density diagram for regression, and the night vision training and testing were completed on the Rgb-T-CC data set. Through verification, the average absolute error of the network is 18.16, the mean square error is 32.14, and the recall rate of target detection is 97.65%. The counting performance and detection performance are superior to the current most advanced bimodal fusion method. The experimental results show that the proposed multimodal feature fusion network can solve the counting and detection problem in night vision environment, and the ablation experiment further proves the effectiveness of parameters of the fusion model.

Key words: crowd counting in night vision environment; Rgb image; thermal imaging; Rgb-T feature fusion

引言

目前热成像设备在监控领域被广泛应用。在夜间及微光特殊条件下，借助热图像和深度学习技术捕获密集人群目标并分析人群数量成为一项具有挑战性的工作[1-2]。在传统的监控设备上嵌入深度学习网络，可在一定程度上解决尺度变化或光噪声问题，但仍然存在一些不足：传统监控设备既无法适应近距离图像目标带来的多尺度的鲁棒性；也无法适应因夜晚光照突降引起的光噪声干扰[3]，目前兼顾这 2 类问题的计数模型较少。

收稿日期：2022-04-11；　**修回日期：**2022-05-30

基金项目：国家自然科学基金（61771266，81571753）；包头市青年创新人才项目（0701011904）

作者简介：任国印（1985—），男，博士研究生，讲师，主要从事深度学习与图像处理方面的研究。E-mail: renguoyin@imust.edu.cn

通信作者：吕晓琪（1963 年），男，博士，教授，主要从事智能图像处理方面的研究。E-mail: lan_tian12345@hotmail.com

一些文献使用摄像头实时观察地面人群聚集的危险情况, 从而实现疏散人群的目的。Miao 在文献 [4] 中提出了一种轻量级卷积神经网络(convolutional neural network, CNN), 该模型能够实时监控地面情况, 以保障人群聚集的安全性。基于深度 CNN 的人群计数方法与浅层学习方法相比, 表现出了明显的性能提升。针对行人尺度变化大的情况, Liu 等人[5] 提出了将多个层中提取的自适应特征结合以生成最终的密度图。Zhou 等人[6] 提出了尺度聚合网络, 改进了多尺度表示, 生成高分辨率密度图。

由于在监控中获取目标时存在角度畸变, 因此倾斜视角获取的人群图像存在视角失真和尺度变化的影响。除此之外高度密集的人群还会受到严重遮挡的影响, 为了解决这个问题, Boominathan 提出了一种新的学习框架 CrowdNet[7], 该框架可从高密度人群中检测到人群目标。该模型由深层和浅层卷积神经网络构成, 并采用了一种基于多尺度金字塔表示的数据增强技术, 该模型对尺度变化具有鲁棒性。然而, 这些方法在夜间图像中会失效。

Samuel 等人[8] 解决了从夜视图像中计算人数的问题, 该方法使用了基于热成像和核密度估计的特征, 比基于 CNN 的方法在夜间人群计数上更准确、更有效。该方法采用快速特征检测器计算密度图, 能降低倾斜视角带来的遮挡问题, 且人群计数的错误率低。然而, 多列 CNN 结构导致训练困难, 冗余参数多, 虽然使用 CNN 的集合可以带来显著的性能提高, 但计算量大。

考虑到上述缺点, 多模态数据融合结合自适应高斯核的方法被用来替代上述方案。一些基于两模态融合的方法被提出[9-11], 通过获得融合特征展示人群计数在夜视光照、遮挡和尺度变换方面的优势。两流模型[12-13] 提出融合层次交叉模态特征, 实现全局代表性的共享特征。此外, 还有一些方法[14] 探讨了共享分支的使用, 将共享信息映射到公共特征空间。

基于以上考虑, 发现热成像对光噪声具有很强的鲁棒性, 并且能够探测到较远的感知距离。Liu[15] 提出了一个多模态的动态增强机制, 可以充分利用鲁棒性更强的热模态图像增加人群计数特征多样性。多模态学习在计算机视觉领域受到越来越多的关注, 通过整合 Rgb 和热成像数据, 该模型引入模态间的软跨模态一致性和最优查询学习来提高鲁棒性。

1　本文所用方法

1.1　热目标自适应高斯核密度图

自适应高斯核能够使得密度图回归更加清晰, 产生与真实度密度图更接近的回归密度图。该自适应高斯核可更接近标注热目标真实头部尺寸, 借助回归产生的密度图可为深度网络提供头部检测先验知识, 引导产生检测框的位置和大小。

人头 Rgb 图像的标签中心点用标准高斯核函数进行计算转换为人群密度图。假设 $C=\{x_1, x_2,\cdots, x_n\}$是 d 维空间上的数据集, 假设一幅图像有 n 个热目标人头实例, 实例量为 n, 那么数据的分布密度可以表示为

$$\hat{f}(x)=\frac{1}{nh^d}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) \tag{1}$$

多元高斯核函数由 (2) 式给出:

$$K\left(\frac{x-x_i}{h}\right)=\frac{1}{2\pi^{\frac{d}{2}}}\cdot\exp\left(-\frac{\|x-x_i\|}{2h^2}\right) \tag{2}$$

式中: x 和 x_i 采用欧式距离计算相似度度量; h 为带宽; d 为维度。当带宽 h 等于对应 Rgb 头部直径 h 时, 估计数据量 n 可以表示为

$$n=\frac{4}{h^{(d+4)}(d+2)} \tag{3}$$

当带宽 h 等于热图像对应头部直径 h_{thermal} 时:

$$n=\frac{4}{h_{\text{thermal}}{}^{(d+4)}(d+2)} \tag{4}$$

对具有多模态数据集 $X=\{x_1,x_2,\cdots,x_n\}$, x_n 代表每个实例, 设其类别标签集 $F=\{c_1,c_2,\cdots,c_f\}$, 其中类别 $c_i(i=1,\cdots,f)$ 中的实例个数为 N_{ci}, 则实例 x_i 关于类别 c_i 的密度计算为

$$f_{ci}(x_i)=\frac{1}{N_{ci}-1}\sum_{i\neq j,i=1}^{n}\frac{1}{R^d_{\text{Rgb}}\left(R^d_{\text{thermal}}\right)}\cdot L(x_i)\cdot K\left(\frac{x_i-x_j}{h}\right),L(x_j)=c_j \tag{5}$$

式中 $L(x_i)$ 和 $L(x_j)$ 分别表示实例 x_i 和 x_j 的标签。该自适应高斯核公式对热图目标适用。

1.2　热成像融合感知器网络设计

热成像与可见光图像的特征融合是一个跨模态问题, 特征数据可能具有较高的模态内可变性, 这使得跨模态图像特征融合任务极具挑战性。

早期融合　早期融合又被称为特征级融合, 原理是在训练过程之前分别提取跨模态图像的区别特征并完成特征级别的融合, 早期融合能充分了

解各个输入模式之间高层次的相互依赖性。然而，这将增加网络的参数，并可能降低模型参数生成的效率。

中期融合　中期融合又被称为中期集成。原理是先分别对每个模态进行一定程度的训练，然后抽取 2 个分支的模态特征并实现中期特征融合。中期融合两流模型可训练模式之间部分高层次的相互依赖特征，又忽略了模态部分共享特征学习，这样做既可减少参数，又能提高融合效率，如图 1 所示。

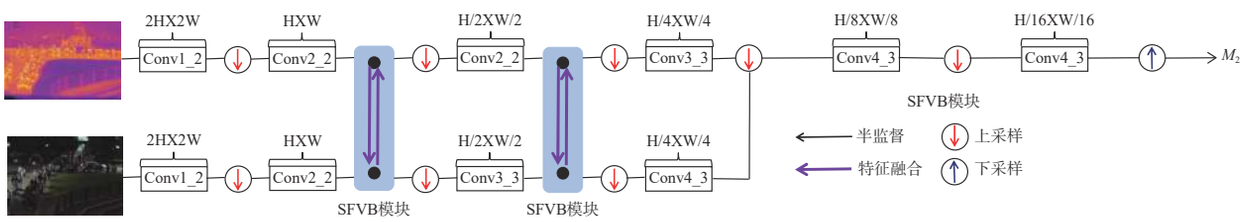


图 1　Rgb-T-net 热图融合感知器结构图

Fig. 1　Structure diagram of Rgb-T-net heatmap fusion perceptron

晚期融合　晚期融合又被称为决策级融合。原理是分别对每个模态进行训练，训练完成后对模型输出层的分类预测分数进行集成以生成最终决策。因此它不允许网络了解各个输入模式之间的这种高层次的相互依赖性。

对于热图分支，我们利用标注热图数据集加强可见光图像特征，完成初始化训练。热图感知器由 2 个分支模块组成，热图特征分支由 6 个卷积层、5 个下采样层和 1 个上采样层组成，可见光特征分支由 4 个卷积层和 3 下采样层组成，卷积层的核尺寸如图 1 所示。Rgb-T-net 模型被用于实现热图分支特征与可见光分支特征的多模态特征融合。

1.3　密度图引导检测

基于检测的模型无法检测到小/微小的头部目标，因为检测子网无法自适应调整这些头部的锚。然而，我们的网络受益于热图自适应高斯核密度图。密度图显示头部分布与密度图中高斯核的位置像素有关。因此，我们将估计的密度图反馈到检测网络中，以提高小型/微型头部的检测的性能。根据网络学习反馈的头部热图像的解码层检测不同尺度的头部。在热图像预处理中，头部热图像被下采样到与密度图相同的大小。利用 $\boldsymbol{M}_l$ 中的每个头部热图像的像素值来加强学习我们估计的密度图。具体地，对于给定训练头部热图像尺寸时，假设要检测的头部大小为标注矩形框尺寸。我们通过训练生成一个头部框特征矩阵 $\boldsymbol{M}_l$，通过 $\boldsymbol{M}_l$ 与密度图函数 D^A 进一步融合生成热图像自适应高斯核约束的密度图：

$$D_l^A(x) = D^A(x) \otimes \boldsymbol{M}_l \tag{6}$$

式中：⊗表示特征融合；$D_l^A(x)$ 是热图像自适应高斯核约束的密度图回归；$\boldsymbol{M}_l$ 是头部框特征矩阵；$D^A(x)$ 是自适应高斯密度图函数。

1.4　系统设计

系统设计流程如图 2 所示。根据热成像和可见光图像特征构造热图融合感知网络，并设置多元高斯函数的初始参数均值 μ 为 0.5，标准偏差 σ 为 0.02。高斯多元函数根据热图标注的中心用高斯核函数带宽逼近热图像头部尺寸。当高斯核函数带宽 h 等于热图像头部尺寸 h_{thermal} 时，停止高斯核扩充并生成高斯核密度图，否则继续匹配头部尺寸。如图 2 所示，本网络设计 3 种融合模式，即早期融合、中期融合和晚期融合。分别对 3 种融合方案进行单独实验验证，因此网络运行前需要预先设定当前融合模式的种类。当融合模式选定后通过 $\boldsymbol{M}_l$ 与密度图函数 D^A 进一步融合生成热

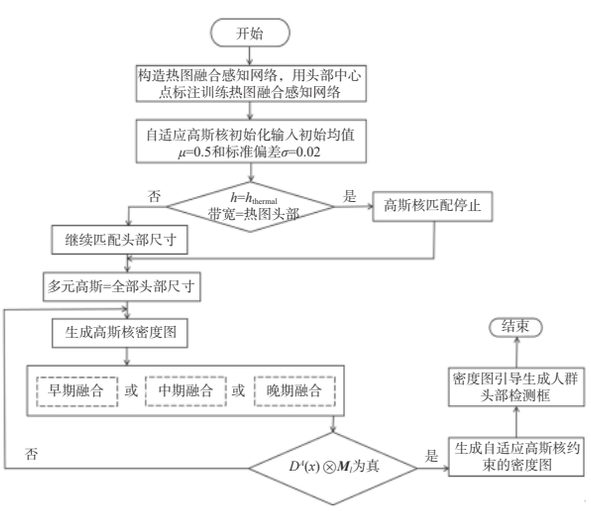


图 2　系统整体设计流程图

Fig. 2　Flow chart of overall system design

图像自适应高斯核约束的密度图并完成计数工作, 最后借助自适应高斯核约束的密度图生成人头检测框。

1.5　多模态融合损失函数

热成像多层感知器结合自适应高斯核可以解决夜视环境头部尺寸自适应感知的问题。其中热成像感知器分别由 Rgb 图像和热图像双模态图像特征融合网络构成, 该网络将多模态特征通过模型中间层导入的方式进行中期融合, 实现特征对齐和自适应融合。

本文提出一种完成 Rgb 特征与热成像特征的有效融合, 从而提高物体识别的准确率。 如图 1 所示, Rgb 图像和 Rgb-T 图像同时作为 Rgb-T-net 网络的输入, 2 个卷积神经网络模型分支分别进行特征学习并分别输出 Rgb 图像和 Rgb-T 图像对应的特征图。融合点 SFVB 选用中期融合模式将独立学习的特征进行融合, 并在融合点对特征图进行融合, 得到融合后的特征图。将 Rgb 图像和 Rgb-T 图像的特征图转换为数据向量, 数据向量在融合函数的作用下生成对应参数矩阵, 并将 2 个卷积神经网络模型连接在一起。融合函数保留了 2 个数据向量的结果, 并将融合后特征图的通道数变为原始特征图的通道数之和, 融合函数定义为

$$f:\boldsymbol{R}_t^a+\boldsymbol{T}_t^b\rightarrow\boldsymbol{A}^{a+b}{}_t \tag{7}$$

式中: $\boldsymbol{R}_t^a$ 和 $\boldsymbol{T}_t^b$ 表示 t 时刻的 Rgb 图像与 Rgb-T 图像分别经过卷积运算得到的空间特征图对应的数据向量; a、b 是原始图像的通道数; $\boldsymbol{A}_t$ 表示融合空间特征图对应的参数矩阵。则与输入数据相对应的参数矩阵 $\boldsymbol{A}$ 可表示为

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1r_n} & \boldsymbol{A}_{1(r_n+1)} & \cdots & \boldsymbol{A}_{1(r_n+r_\tau)}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2r_n} & \boldsymbol{A}_{2(r_n+1)} & \cdots & \boldsymbol{A}_{2(r_n+r_r)}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \boldsymbol{A}_{k1} & \boldsymbol{A}_{k2} & \cdots & \boldsymbol{A}_{kr_n} & \boldsymbol{A}_{k(r_n+1)} & \cdots & \boldsymbol{A}_{k(r_n+r_r)}\end{bmatrix} \tag{8}$$

$\boldsymbol{A}$ 的前半部分是与 Rgb 图像向量相对应的参数, 后一部分是与 Rgb-T 图像向量相对应的参数:

$$\boldsymbol{A}_{\text{Rgb}}=\begin{bmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1r_n}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2r_n}\\ \vdots & \vdots & \vdots & \vdots\\ \boldsymbol{A}_{k1} & \boldsymbol{A}_{k2} & \cdots & \boldsymbol{A}_{kr_n}\end{bmatrix} \tag{9}$$

$$\boldsymbol{A}_{\text{thermal}}=\begin{bmatrix}\boldsymbol{A}_{1(r_n+1)} & \cdots & \boldsymbol{A}_{1(r_n+r_r)}\\ \boldsymbol{A}_{2(r_n+1)} & \cdots & \boldsymbol{A}_{2(r_n+r_r)}\\ \vdots & \vdots & \vdots\\ \boldsymbol{A}_{k(r_n+1)} & \cdots & \boldsymbol{A}_{k(r_n+r_r)}\end{bmatrix} \tag{10}$$

λ 是与热图像向量对应的参数, k 表示所有可能的类别标签:

$$\lambda_{\text{Rgb}}=\begin{bmatrix}\lambda_{r1} & & & \\ & \lambda_{r2} & & \\ & & \ddots & \\ & & & \lambda_{rk}\end{bmatrix} \tag{11}$$

$$\lambda_{\text{thermal}}=\begin{bmatrix}\lambda_{i1} & & & \\ & \lambda_{i2} & & \\ & & \ddots & \\ & & & \lambda_{ik}\end{bmatrix} \tag{12}$$

整体代价函数如(13)式所示:

$$\begin{aligned}J_{\text{sparse}}=&\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{w,b}\left(x^i\right)-x^i\right\|^2\right)+\\&\frac{1}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l^r}\sum_{j=1}^{s_{l+1}^r}\left(\lambda_{\text{Rgb}}\left(A_{\text{Rgb}}^l\right)_{ij}\right)^2+\\&\left(\frac{1}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l^l}\sum_{j=1}^{s_{l+1}^l}\left(\lambda_{\text{thermal}}\left(A_{\text{thermal}}^l\right)_{ij}\right)^2+\beta P(x)\right)\end{aligned} \tag{13}$$

物体 Rgb 特征对应的权重衰减参数 λ_{Rgb} 初始化一个较小的值, 减小其惩罚力度, 提取更多的 Rgb 特征。物体热成像特征对应的权重衰减参数 λ_{thermal} 初始化一个较大的值, 加大其惩罚力度, 提取更少的热图特征。

2　模型训练与实验评价

2.1　数据集介绍与评价标准

1) 数据集介绍

热图感知器在 Rgb-T-CC 数据集上完成了实验评估, 通过实验对比验证我们提出方法的可行性与适用性。实验所用关键数据集的参数如表 1 所示。在 Rgb-T-CC 数据集下本文方法与目前该数据集下最先进的人群计数方法做了结果对比, 并给出每个数据集图像上的密度图及人群真实值和估计值。同时给出消融实验结果, 用以证明我们的方法综合体中每个方法单元的独立有效性, 最后给出夜视人群计数和检测效果图。

表 1　Rgb-T-CC 数据集的参数信息

Table 1　Parameter information for Rgb-T-CC dataset

数据集	分辨率	数据类型	数量	最大	最小	平均	总计	模态
Rgb-T-CC	640×480	Rgb+T	4060	82	45	68	138,389	Rgb-T

本文采用人群计数工作中的常用评价标准，使用平均绝对误差（MAE）和均方误差（MSE）来评估不同的方法：

$$\mathrm{MAE}=\frac{1}{N}\sum_{1}^{N}\left|z_i-z_j\right| \tag{14}$$

$$\mathrm{MSE}=\sqrt{\frac{1}{N}\sum_{1}^{N}(z_i-z_j)^2} \tag{15}$$

式中：N 是测试图像的总数量；z_i 是第 i 幅测试图像中的实际人数；z_j 是第 i 幅图像中的人数估计。

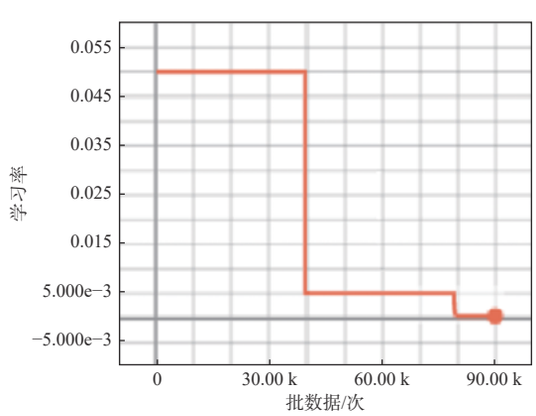
2）数据集参数设置和训练

Rgb-T-CC 是基准 Rgb-T 数据集[15]，它包括 Bright 和 Dark 两个部分图像。Bright 训练集包括 510 对图像，验证集包括 97 对图像，测试集包括 406 对图像；Dark 训练集包括 520 对图像，验证集包括 103 对图像，测试集包括 394 对图像；这些图像被随机分到最终的 3 个新的训练集、验证集和测试集。最终 1 030 对图像用于训练，200 对图像用于验证，800 对图像进行测试。

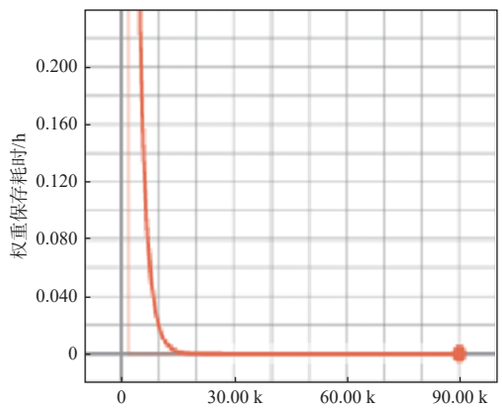
在训练前需要设置训练模型的相应参数。我们端到端地训练 Rgb-T-net 模型。自适应高斯核的高斯参数 μ 由平均值设置为 0.5，标准偏差 σ 设置为 0.02。在实验中，Rgb-T-net 选择带动量的随机梯度下降 SGD（stochastic gradient descent），初始学习率设置为 0.005，动量设置为 0.85。训练过程如图 3 所示。我们的方法均在 Pytorch 框架下实现，硬件方面使用了 3 块 NVIDIA 3080 Ti GPU 显卡和 4 块 Intel(R)E5-2630 v4 CPU 以确保显卡和运算单元的性能需要。

2.2　与目前最先进的方法比较

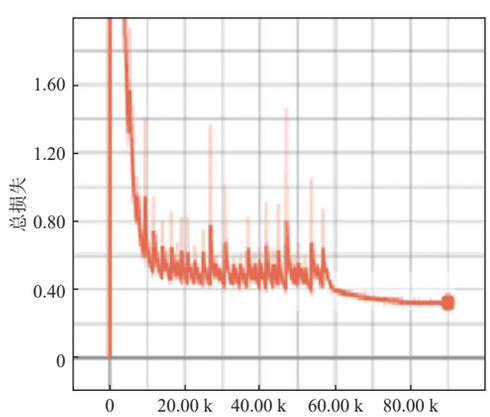
Rgb-T-CC 数据集是中山大学公布的热成像人群计数数据集。如表 2 所示，给出本文所用方法与目前先进的人群计数方法在 Rgb-T-CC 数据集下的对比结果。由于目前中山大学团队的实验新方法是第一个使用 Rgb-T-CC 数据集并完成不同融合模型的比较。我们的方法在这个方案的基础上做了模型和融合策略改进。仍使用 MCNN[15]、SANet[15]、CSRNet[16] 和 Bayesian Loss[17] 等经典计数



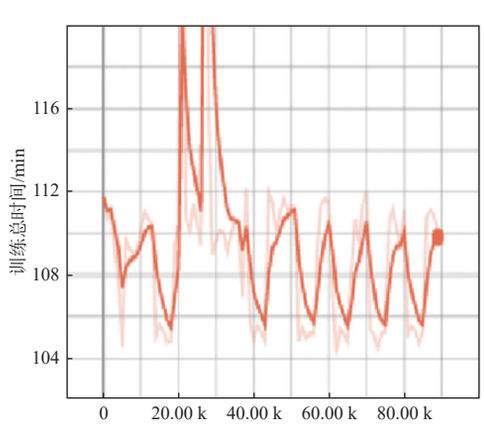
(a) 学习率设置曲线



(b) 权重的变化曲线



(c) 损失函数变化曲线



(d) 训练总时间变化曲线

图 3　模型训练过程的曲线参数可视化

Fig. 3　Visualization of curve parameters in model training process

表 2　Rgb-T-CC 数据集上不同最新方法的比较

Table 2　Comparison of different state-of-the-art methods on Rgb-T-CC dataset

模型(Rgb-T-CC数据集)	MAE	MSE
UCNet[18]	33.96	56.31
HDFNet[19]	22.36	33.93
BBSNet[20]	19.56	32.48
MCNN[15]	21.89	37.44
CSRNet[16]	20.4	35.26
Bayesian Loss[17]	18.7	32.67
MCNN+IADM[15]	19.77	30.34
CSRNet+IADM[15]	17.94	30.91
Bayesian Loss+IADM[15]	15.61	28.18
本文Rgb-T-net	18.16	32.14

模型作为主干网进行实验参照, 与结果最好的 Bayesian Loss 进行对比后发现, Rgb-T-net 模型在 Rgb-T-CC 数据集上 MAE、MSE 的误差有 0.54、0.53 的提升。同样本文方法与 UCNet[18]、HDFNet[19] 和 BBSNet[20] 多个表现最好的多模模态融合模型进行对比。与结果最好的 BBSNet 对比后发现, Rgb-T-net 模型在 Rgb-T-CC 数据集上 MAE、MSE 的误差有 1.4、0.34 的提升。中山大学团队的文献中将 MCNN、SANet[21]、CSRNet 和 Bayesian Loss 等经典计数模型在网络中整合 IADM“早期融合”机制可以提升模型性能, 最明显的表现分别是 MCNN+IADM 的 MAE、MSE 的误差有 2.12、2.14 的提升; SANet+IADM 的 MAE、MSE 的误差有 3.81、7.88 的提升; CSRNet+IADM 的 MAE、MSE 的误差有 2.56、4.35 的提升; Bayesian Loss+IADM 的 MAE、MSE 的误差有 3.09、4.49 的提升。不同的是, 我们没有使用“早期融合”的方式将 Rgb 和热成像特征作为输入, 而是提出一种全新的网络并采用“中期融合”方式与中山大学的“早期融合”结果做一个对比。除此之外, 本文使用自适应高斯核完成密度图回归。进行对比后发现本文“中期融合”模型和中山大学的“早期融合”方案在 Rgb-T-CC 数据集上的 MAE、MSE 误差有进一步的提升。

我们证实了该方法在不同光照条件下借助热图生成密度图方面的有效性, 在 Rgb-T-CC 数据集上挑选了不同光照条件的图像作为计数对象。如图 4 所示, 第 1 列图像从上到下的光照条件逐渐变暗, 最明显的是第 1 张图像是白天室内场景, 最下面一张图像是夜间街景。第 2 列图像是不同光照条件下的热图像。第 3 列图像是 Bayesian Loss+IADM 在不同光照条件下的人群密度图。第 4 列图像是 Bayesian Loss+RDNA 在不同光照条件下的人群密度图。

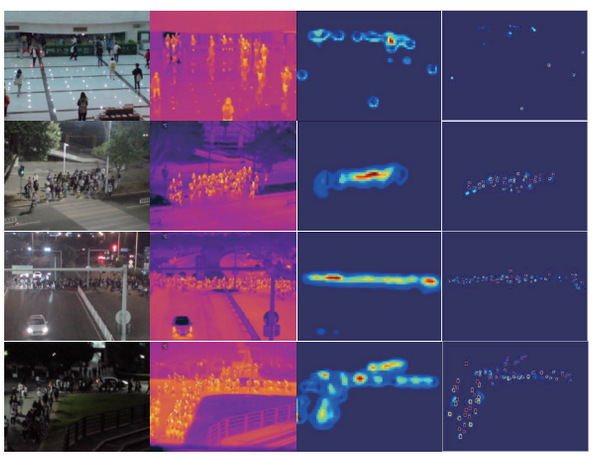


图 4　ShanghaiTechPartA, ShanghaiTechPartB, UCF-QNRF 和 UCF_CC_50 数据集的可视化结果。从左到右: 输入图像、真实标注、贝叶斯结果和我们推荐方法的结果

Fig. 4　Visualization results from ShanghaiTechPartA, ShanghaiTechPartB, UCF-QNRF and UCF_CC_50 datasets (from left to right: input images, real annotations, Bayesian results and results from proposed method)

从 Rgb-T-CC 数据集上生成的密度图结果来看, 本文方法密度图比 Bayesian Loss+RDNA 的密度图人群散布可分性更好, 这也一定程度上证明了热图感知器的有效性和计数性能。当人群光照变得非常暗时(如场景 4 的夜间黑暗光照条件下)热图的融合特性起到有效作用, 没有受到光噪声影响, 人群散布可分性依然明显, 因此证明了本方法在不同光照下的有效性。

2.3　人群定位

目前, 很难做到在遮挡明显环境下的密集人群人头检测, 尤其是夜间检测往往会失效。为了解决遮挡问题和夜间人头检测对小像素人头的有效性, 本文使用热量图感知器完成夜间人头目标检测。借助自适应高斯核可完成回归引导检测的人头计数和检测, 如图 5 所示。图 5 是基于热图感知器+自适应高斯核实现密度回归引导人头检测的结果, 第 1 张图来自 Rgb-T-CC 的人群热成像, 第 2 张图是本文方法生成的估计检测框。本文所用

方法是将数据集标注框的中心作为训练目标根据自适应高斯核与对应感知器定位估计框的边界。我们从真实标注中提取头部位置点，利用热图感知器可以检测到夜间环境中的人头，在黑暗图像下方遮挡人群头部仍然可被检测。

(a) Rgb-T人群热成像　　(b) Rgb-T人群检测

图 5　热图数据集上的人群检测结果

Fig. 5　Crowd detection results on heatmap datasets

3　融合有效性消融实验

3.1　融合位置的有效性分析

本文对热图感知器的有效性进行了消融实验研究。如表 3 所示，选择了 4 种不同的变量来定性分析，分别使用早期融合、中期融合和晚期融合的方式将 Rgb 和热成像特征构造热图感知网络，不同阶段融合的网络会出现明显的差异，当仅使用了自适应高斯核 AGK（adaptive gaussian kernel）的误差 MAE=22.46、MSE=38.97。与仅使用 AGK 完成密度图回归相比，热成像感知网络可以约束每个高斯核的边缘扩充，对密集人群更有效。这意味着 Rgb-T 热成像感知器+AGK 的组合有助于人群计数，MAE 和 MSE 误差明显减小，其中早期融合 AGK+Rgb-T-net 在 MICC 数据集的 MAE=18.01、MSE=31.49；中期融合 AGK+Rgb-T-net 在 MICC 数据集的 MAE=18.16、MSE=32.14；晚期融合 AGK+Rgb-T-net 在 MICC 数据集的 MAE=19.35、MSE=34.71。因此本文所用热图感知器的中期融合+AGK 的方法可满足时间和训练数据缺少的前提，误差精度明显提升。

表 3　Rgb-T-CC 数据集不同融合方式的比较

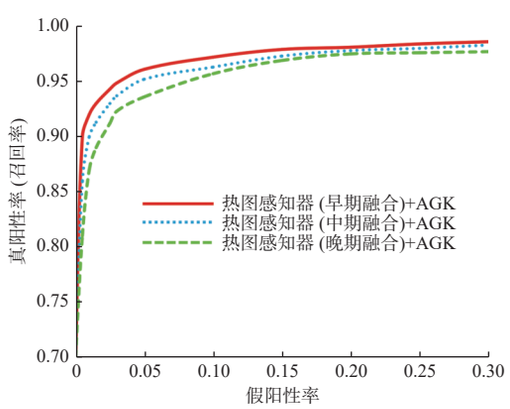
Table 3　Comparison of different fusion methods on Rgb-T-CC dataset

融合方式(Rgb-T-CC数据集)	MAE	MSE
AGK	22.46	38.97
AGK+Rgb-T-net(早期融合)	18.01	31.49
AGK+Rgb-T-net(中期融合)	18.16	32.14
AGK+Rgb-T-net(晚期融合)	19.35	34.71

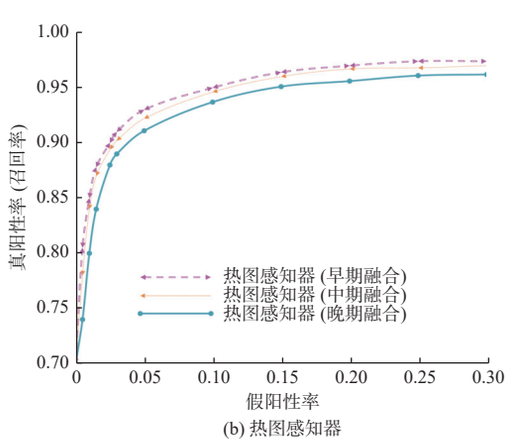
3.2　检测模型的有效性分析

不同的融合网络和检测模型在人群检测结果的召回率上有较大的差异性，如图 6 所示。Rgb 分支和热图像分支的早期融合理论上比中级融合更能充分了解各个输入模式之间高层次的相互依赖性和精度优势，如图 6(a) 和如图 6(b) 所示。在热图感知器+AGK 和热图感知器 2 种模型中，早期融合的召回率大于中期融合的召回率，中期融合的召回率大于晚期融合的召回率。自适应高斯核函数 AGK 可提高热图感知器的密度图精度，因此热图感知器+AGK 对应召回率整体上大于热图感知器，如图 6(c) 所示。然而，更高的精度优势带来的代价是训练可能需要更多的数据，但目前 Rgb-T 数据集并没有 Rgb 数据集多，因此早期融合受到数据量的限制提升有限。

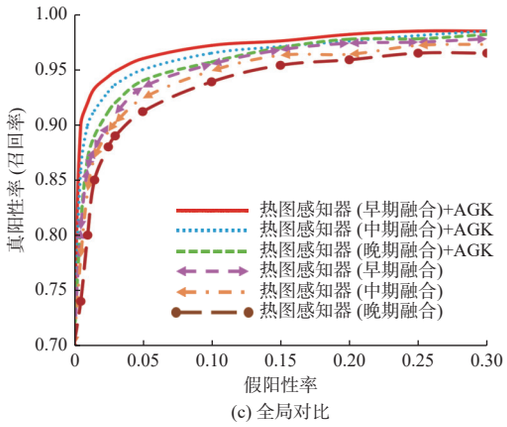
使用中期融合首先训练 Rgb-T 分支的权重，Rgb-T 深度分支可以在 Rgb 网络的一个最大池层之前合并为 1×1 卷积层。这样做的目的在于 Rgb-T 分支可以使用已经在 Rgb 分支上预先训练的权重，且网络可充分了解各个输入模式之间高层次的相互依赖性。这将大大降低总训练时间，而且这样不需要更多的 Rgb-T 训练数据，这也满足 Rgb-T



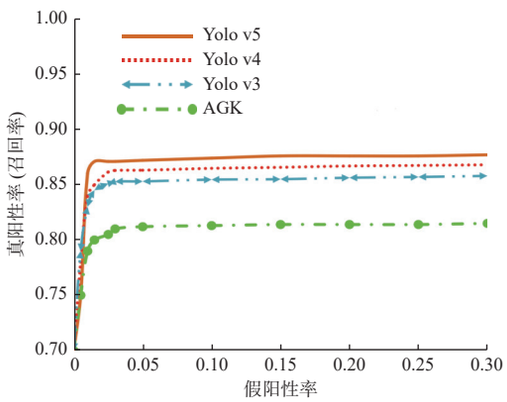
(a) 热图感知器+AGK



(b) 热图感知器



(c) 全局对比



(d) 其他检测模型

图 6　平均精度召回曲线对比

Fig. 6　Comparison of average precision recall curves

数据集不足的实际问题, 虽然在最终精度方面中期融合没有早期融合的误差小, 但相差只有不到1% 的召回率误差, 且 MAE、MSE 误差仅相差 MAE=0.021、MSE=0.032, 牺牲的精度与训练更多的数据、消耗更多时间相比可以被接受。

Rgb 网络和 Rgb-T 深度分支晚期融合的优势是不必反复初始化网络权重, 尽管有额外的输入网络权重仍可重用, 但它不允许网络了解各个输入模式之间的这种高层次的相互依赖性, 只给出了分类级别上的分数融合结果。

图 6(d) 给出了目前人群检测领域常用的 Yolo v3、Yolo v4、Yolo v5 以及单独使用自适应高斯核 AGK 的召回率, 从图中可以看出, 仅用自适应高斯核 AGK 没有热图感知器的夜间人群检测召回率性能明显降低, 且没有经典的 Yolo v 系列检测模型的召回率精度高。

4　结论

提出了一种用于夜视环境下的拥挤人群计数的热图融合网络(Rgb-T-net)及结合 AGK 的引导检测方法, 建立了一种 Rgb 和热图像的跨模态融合的人群计数和密度估计方法。在 Rgb-T-CC 公开数据集上进行了大量的实验和评估, 该方法在训练时间和检索召回率方面优于现有多模态融合的人群计数方法, 同时对 Rgb 人群计数任务有很好的推广作用。使用中期融合的网络可提取图像通道特征, 使用高斯模型可提取图像的空间边缘约束特征, 用于最终的人群计数估计。本文方法在人群计数方面取得了满意的结果。在引导检测方面, Rgb-T-net+AGK 可实现人群的夜视计数与检测。从结果来看, 本文的方法平均绝对误差为 18.16, 均方误差为 32.14, 检测召回率为 97.65%, 对遮挡和夜间复杂场景鲁棒性较好。在未来, 将把本文提出的方法扩展到视频人群计数与检测中, 提升整体算法的实时处理能力。

参考文献:

[1] XU Mingliang, GE Zhaoyang, JIANG Xiaoheng, et al. Depth Information Guided Crowd Counting for complex crowd scenes[J]. Pattern Recognition Letters, 2019, 12(5): 563-569.

[2] 高凯珺, 孙韶媛, 姚广顺, 等. 基于深度学习的无人车夜视图像语义分割[J]. 应用光学, 2017, 38(3): 421-428.
GAO Kaijun, SUN Shaoyuan, YAO Guangshun, et al. Semantic segmentation of night vision images for unmanned vehicles based on deep learning[J]. Journal of Applied Optics, 2017, 38(3): 421-428.

[3] 吴海兵, 陶声祥, 张良, 等. 低照度条件下三基色获取及真彩色融合方法研究[J]. 应用光学, 2016, 37(5): 673-679.
WU Haibing, TAO Shengxiang, ZHANG Liang, et al. Tricolor acquisition and true color images fusion method under low illumination condition[J]. Journal of Applied Optics, 2016, 37(5): 673-679.

[4] MIAO Yunqi, HAN Jungong, GAO Yongsheng, et al. ST-CNN: spatial-temporal convolutional neural network for crowd counting in videos[J]. Pattern Recognition Letters, 2019, 125(3): 113-118.

[5] LIU X, YANG J, DING W, et al. Adaptive mixture regression network with local counting map for crowd counting[C]//European Conference on Computer Vision, August 23-28, 2020, Glasgow. UK: Springer, 2020: 241-257.

[6] ZHOU Yuan, YANG Jianxing, LI Hongru, et al. Ad-

versarial learning for multiscale crowd counting under complex scenes[J]. IEEE Transactions on Cybernetics, 2021, 51(11): 5423-5432.

[7] BOOMINATHAN L, KRUTHIVENTI S S S, BABU R V. Crowdnet: a deep convolutional network for dense crowd counting[C]//Proceedings of the 24th ACM international conference on Multimedia, October 15-19, 2016, New York, NY. United States: ACM, 2016: 640-644.

[8] SAMUEL M, SAMUEL-SOMA M A, MOVEH F F. Ai driven thermal people counting for smart window facade using portable low-cost miniature thermal imaging sensors[J]. 2020, 16(5): 1566-1574.

[9] LIU D, ZHANG K, CHEN Z. Attentive cross-modal fusion network for RGB-D saliency detection[J]. IEEE Transactions on Multimedia, 2020, 23(1): 967-981.

[10] XU G, LI X, ZHANG X, et al. Loop closure detection in rgb-d slam by utilizing siamese convnet features[J]. Applied Sciences, 2022, 12(1): 62-75.

[11] TANG Z, XU T, LI H, et al. Exploring fusion strategies for accurate rgbt visual object tracking[J]. ArXiv Preprint ArXiv: 2201.08673, 2022.

[12] ZHANG W, GUO X, WANG J, et al. Asymmetric adaptive fusion in a two-stream network for RGB-D human detection[J]. Sensors, 2021, 21(3): 916-921.

[13] ZHOU Wujie, JIN Jianhui, LEI Jingsheng, et al. CEGFNet: common extraction and gate fusion network for scene parsing of remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 19(6): 1524-1535.

[14] ZHANG Shihui, LI He, KONG Weihang. A cross-modal fusion based approach with scale-aware deep representation for RGB-D crowd counting and density estimation[J]. Expert Systems with Applications, 2021, 180(5): 115071.

[15] LIU Lingbo, CHEN Jiaqi, WU Hefeng, et al. Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, Long Beach, CA. USA: IEEE, 2021: 4823-4833.

[16] LI Yuhong, ZHANG Xiaofan, CHEN Deming. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, June 18-22, 2018, Salt Lake City, Utah. USA: IEEE, 2018: 1091-1100.

[17] FISCHER M, VIGNES A. An imprecise bayesian approach to thermal runaway probability[C]//International Symposium on Imprecise Probability: Theories and Applications, July 6-9, 2021, University of Granada, Granada. Spain: PMLR, 2021: 150-160.

[18] LIU Lingbo, CHEN Jiaqi, WU Hefeng, et al. Cross-modal collaborative representation learning and a large-scale RGBT benchmark for crowd counting[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 20-25, 2021. Nashville, TN. USA: IEEE, 2021: 4823-4833.

[19] LIU Z, HE Z, WANG L, et al. Visdrone-cc2021: The vision meets drone crowd counting challenge results[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, October 11- October 17, 2021, Montreal, BC, Canada. USA: IEEE, 2021: 2830-2838.

[20] ZHOU Wujie, GUO Qinling, LEI Jingsheng, et al. ECFFNet: effective and consistent feature fusion network for RGB-T salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 1224-1235.

[21] FAN J, YANG X, LU R, et al. Design and implementation of intelligent inspection and alarm flight system for epidemic prevention[J]. Drones, 2021, 5(3): 68-82.